LAD: Learnable Anisotropic Deformation for 3D Medical Image Analysis

Chengcai Liu[1*], Haolin Sang[1*], Xingyu Huang[1], Yi Wu[1], Chenghao Wang[1], Weiqiang Wu[1], Guangjian Zeng[3], Jie Tian[1,2], and Shuo Wang[1†]

[1] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing, China
[2] CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] Southern University of Science and Technology, Shenzhen, China

Abstract. Three-dimensional medical images inherently exhibit spatial anisotropy due to non-uniform slice spacing and heterogeneous acquisition protocols. However, most 3D Transformer architectures assume isotropic voxel grids and rely on fixed resampling to standardize inputs. This mismatch introduces two key limitations: (1) forced resampling distorts native anatomical geometry and may degrade fine structural details; (2) isotropic positional encoding may overlook instance-specific spacing, limiting spatial precision. While recent anisotropy-aware designs introduce direction-specific priors, they typically depend on static, dataset-level scaling factors that require manual hyperparameter tuning and lack instance-specific adaptivity. To address these limitations, we propose Learnable Anisotropic Deformation (LAD), a geometry-aware adaptation framework for 3D Transformers. LAD learns instance-specific spatial scaling factors directly from image features and incorporates them into both positional encoding and feature modulation. We design a frequency-modulated anisotropic Rotary Positional Embedding (RoPE) to adjust spatial phase and an adaptive anisotropic feature modulation module to recalibrate semantic responses along biased axes. This design enables Transformers to respect native volumetric structure without requiring explicit spacing metadata. We integrate LAD into three representative 3D backbones and evaluate it on volumetric segmentation and multi-disease classification benchmarks. Across four datasets (17,359 3D volumes) and six tasks spanning CT and MRI, LAD consistently improves performance across backbones. Further analysis shows that the learned scales correlate with physical spacing in segmentation and adapt across classification tasks, supporting both geometric consistency and task adaptivity. The code is available at https://github.com/chengcailiu/LAD.

Keywords: Learnable Anisotropic Deformation · Vision Transformers · Positional Encoding · 3D Medical Image Analysis · Spatial Bias.

* These authors contributed equally to this work.
† Corresponding author: shuo_wang@buaa.edu.cn.

1 Introduction

Transformers have revolutionized medical image analysis, shifting the paradigm from task-specific CNNs to general-purpose 3D representation learners [4, 6, 10]. By leveraging self-attention, these architectures capture long-range dependencies and global semantic contexts. Recently, scaling these volumetric backbones has yielded robust foundation models that establish a unified methodology for understanding complex human anatomy [1, 2, 15–17].

Despite this versatility, a fundamental conflict persists: the isotropic architectural design of standard Transformers clashes with the inherent *spatial biases* of medical data. These biases stem from both physical acquisition and anatomical morphology [12]. To bridge this gap, prevailing pipelines force data homogenization through fixed spatial resampling [8] or isotropic grids. However, this forced homogenization is fundamentally suboptimal, as irreversible resampling inevitably compromises data fidelity [11]. Rather than irremediably distorting native data to fit rigid architectures, an ideal network should adapt its internal computational geometry to naturally accommodate these composite spatial biases.

Transformers heavily depend on Positional Encodings (PEs) to interpret 3D structures. However, standard Absolute PEs and isotropic Rotary Positional Embeddings (RoPE) [14] often assume a regular voxel grid. While recent direction-specific PEs [9] introduce static priors, adapting to dynamic inter-subject variability remains challenging. Crucially, spatial biases extend beyond positional misalignment, as non-uniform spatial resolutions can also distort local textures within semantic feature representations. Although anatomical priors frequently provide spatial guidance, directly translating dynamic geometric scales into semantic feature modulation remains relatively underexplored.

To concurrently address these positional and semantic limitations, we propose **LAD (Learnable Anisotropic Deformation)**, a versatile architectural framework designed to dynamically adapt to physical spacing and anatomical morphology. Here, "deformation" denotes an internal anisotropic reparameterization of spatial phase and feature amplitude, rather than voxel-level spatial warping. By recalibrating both positional embeddings and feature representations to respect native volumetric geometry, this work makes the following contributions:

- We introduce an **Instance-aware Anisotropy Learner** to dynamically regress spatial scaling factors from image textures, empowering the model to perceive non-uniform spatial resolutions and orientation biases.
- We propose a **Frequency-modulated Anisotropic RoPE** that aligns positional encoding frequencies with the perceived geometric space to maintain consistent spatial attention across directionally biased dimensions.
- We design an **Adaptive Anisotropic Feature Modulation (AAFM)** mechanism that leverages learned geometric priors to residually gate feature channels, acting as a dynamic filter to explicitly recalibrate axis-specific semantic representations.

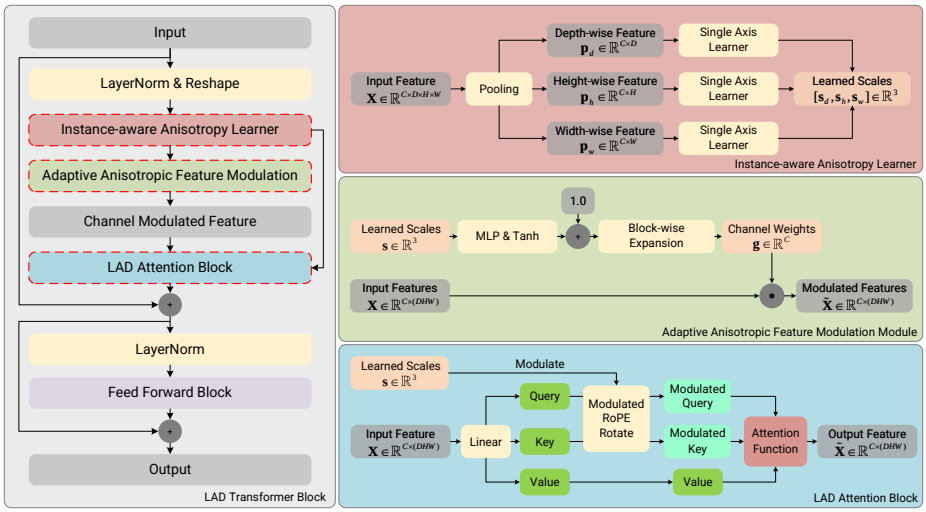

Fig. 1. Overview of the Learnable Anisotropic Deformation (LAD) framework. The *Instance-aware Anisotropy Learner* dynamically extracts physical scaling factors ($\mathbf{s}$) from image features. These geometry priors are bifurcated into two pathways: 1) modulating the rotation frequencies in Anisotropic RoPE (spatial phase), and 2) gating the channel representations via AAFM (feature amplitude), ensuring the Transformer seamlessly adapts to native physical geometries.

2 Methodology

2.1 Overview of the Learnable Anisotropic Deformation Framework

As illustrated in Fig. 1, LAD operates by dynamically perceiving composite spatial biases and injecting them into both the attention mechanism and the feature space. Specifically, given an intermediate feature representation, LAD first employs the **Instance-aware Anisotropy Learner** to regress an explicit geometric scaling factor $\mathbf{s}$. This prior is then bifurcated into two parallel modulation pathways. In the spatial pathway, the **Frequency-modulated Anisotropic RoPE** scales the base rotation frequencies by $\mathbf{s}$ to physically calibrate the self-attention coordinate space. Concurrently, in the semantic pathway, the **AAFM** leverages $\mathbf{s}$ to residually recalibrate the feature channels. This dual-pathway design facilitates geometry-consistent modeling of both spatial dependencies (the "where") and semantic representations (the "what").

2.2 Instance-aware Anisotropy Learner

To automate the perception of geometric priors from both acquisition and anatomy, we design an axis-specific hyper-network. For backbone compatibility, input features are reshaped into their canonical 3D topology $\mathbf{X} \in \mathbb{R}^{C\times D\times H\times W}$ by unflattening patch sequences. Along each axis $k \in \{d, h, w\}$, we summarize transverse

planes via an orthogonal mean pooling $\mathcal{P}_{\perp k}$ (e.g., $\mathcal{P}_{\perp d}$ averages $\{H, W\}$). This 1D structural profile is processed through a 1D Convolution to capture local inter-slice dependencies, followed by Global Average Pooling (GAP) and an MLP to regress the instance-specific scale $\mathbf{s}_k$:

$$\mathbf{s}_k = 1.0 + \tanh\left(\mathrm{MLP}_k\left(\mathrm{GAP}\left(\mathrm{Conv1D}_k\left(\mathcal{P}_{\perp k}(\mathbf{X})\right)\right)\right)\right) \tag{1}$$

Zero-initializing the MLP's final linear layer promotes $\mathbf{s}_k = 1.0$ at training onset, facilitating a stable transition from pre-trained isotropic states. These predicted scales then guide a dual-calibration process: Anisotropic RoPE adjusts the spatial phase of coordinates, while AAFM recalibrates the semantic amplitude of features along the target axes.

2.3 Frequency-modulated Anisotropic RoPE

Standard RoPE encodes spatial distances by rotating token vectors in the complex plane using a fixed frequency set Θ. However, this static approach implicitly assumes an isotropic coordinate space. To bridge this gap, we propose an Anisotropic RoPE that explicitly modulates the base frequencies using our learned geometric priors to match the physical reality.

For a patch token located at the 3D coordinate $\mathbf{p} = (p_d, p_h, p_w)$, anisotropic RoPE is applied independently within each attention head. For an attention layer with A heads, the Anisotropy Learner produces a three-dimensional scale vector $\mathbf{s}^{(a)} = [s_d^{(a)}, s_h^{(a)}, s_w^{(a)}]$ for the a-th head, corresponding to the depth, height, and width axes. Within each head, the channel dimension is equally partitioned into three groups, and each group is assigned to one spatial axis. The modulated rotation frequency for axis $k \in \{d, h, w\}$ in head a is dynamically defined as $\tilde{\Theta}_k^{(a)} = \Theta \cdot s_k^{(a)}$. Consequently, the attention score between a query at position $\mathbf{p}$ and a key at position $\mathbf{q}$ becomes a function of the physically-scaled relative distance:

$$\mathrm{Score}(\mathbf{p}, \mathbf{q}) \propto \sum_{a=1}^{A} \sum_{k \in \{d,h,w\}} \cos\left((p_k - q_k) \cdot \Theta \cdot s_k^{(a)}\right) \tag{2}$$

Intuitively, if the Anisotropy Learner predicts a large scale $\mathbf{s}_d > 1.0$ for a highly down-sampled depth axis, the spatial frequency proportionally increases. A discrete coordinate step of $\Delta z = 1$ thereby produces a significantly larger phase shift, effectively stretching the perceived mathematical distance to compensate for the true physical slice thickness.

2.4 Adaptive Anisotropic Feature Modulation (AAFM)

While anisotropic RoPE recalibrates the spatial coordinate space, non-uniform spatial resolutions also distort the semantic feature representations. To address this, we propose Adaptive Anisotropic Feature Modulation (AAFM) to dynamically modulate channel responses based on the perceived geometry.

Given the learned scaling vector $\mathbf{s} = [\mathbf{s}_d, \mathbf{s}_h, \mathbf{s}_w]$, AAFM generates a channel-wise gating signal $\mathbf{g} \in \mathbb{R}^C$, where C denotes the channel dimension of each attention head. To match the high-dimensional feature space, we first map $\mathbf{s}$ to an axis-specific gate $\Delta\mathbf{g} \in \mathbb{R}^3$ via a lightweight MLP, which is then expanded block-wise and padded to geometrically align with the respective channel groups of Anisotropic RoPE, ensuring consistency between the spatial phase modulation in RoPE and the semantic amplitude modulation in AAFM. The modulated features $\tilde{\mathbf{X}}$ are computed as:

$$\tilde{\mathbf{X}} = \mathbf{X} \odot (1.0 + \tanh(\mathrm{MLP}_{gate}(\mathbf{s}))) = \mathbf{X} + \mathbf{X} \odot \Delta\mathbf{g} \tag{3}$$

By employing a residual formulation with zero-initialized MLP_{gate}, AAFM acts as a dynamic, orientation-aware filter that can selectively amplify or attenuate feature channels without disrupting the original representation space.

3 Experiments

3.1 Backbone Integration

To evaluate architectural generalizability, we integrate LAD (Anisotropy Learner, Anisotropic RoPE, and AAFM) into three representative 3D Transformer backbones: UNETR (ViT), SwinUNETR (Swin Transformer), and Primus (EVA-02). In each case, we integrate anisotropic RoPE into the attention blocks, replacing or augmenting the original positional encoding depending on the backbone. Meanwhile, the Learner and AAFM are embedded within the transformer blocks to facilitate adaptive modulation of spatial and semantic representations.

3.2 Datasets

Volumetric Segmentation. AbdomenAtlas 1.0 mini [13] features 5,195 multi-center CT scans and 9 organs, providing diverse Z-axis spacing to evaluate LAD's perception of physical anisotropy ratios at scale. Following a 4:1 train-test split, we perform 5-fold cross-validation on the training set and report mean Dice (mDice) and mean HD95 (mHD95) results via ensemble inference on the test set.

Downstream Classification. To evaluate the robustness and generalizability of the learned representations, we assess LAD on five downstream classification tasks across heterogeneous imaging modalities and clinical settings.

1. **RadChest-CT** [5] targets thoracic abnormality prediction to evaluate task-adaptive scaling across diverse pathological morphologies within a standardized input grid. We focus on **three abnormality prediction tasks**: atelectasis (T1), lung opacity (T2), and bronchiectasis (T3). The dataset includes 3,630 CT scans with the official training, validation, and test split[4].

2. **UCSF-MRI** [3] involves brain IDH mutation prediction (T4) to evaluate cross-modality stability and LAD's robustness under small-scale data regimes.

[4] https://zenodo.org/records/6406114

The dataset includes 495 patients, and is randomly divided into training, validation, and test sets in a 7:1:2 ratio.

3. **MultiCenter-CT** targets multi-center lung cancer subtype prediction (T5, squamous cell carcinoma vs. adenocarcinoma) to evaluate framework resilience to inter-site heterogeneity and varying acquisition protocols. The dataset includes 6,804 patients with lung cancer from Center 1 (West China Hospital of Sichuan University) and 1,235 patients from Center 2 (The first hospital of Jilin University). Center 1 is randomly divided into training and validation sets with an 8:2 ratio, while Center 2 serves as an independent external test set.

3.3 Implementation Details

Volumetric Segmentation. Using an adapted **nnU-Net**, we disable spatial resampling and restrict augmentations to in-plane transformations to strictly preserve native physical spacings and anatomical orientations. Models are trained with Dice and Cross-Entropy loss, using 96^3 patches, a batch size of 4. The initial learning rates for UNETR, SwinUNETR, and Primus are configured following the standard baselines [15].

Downstream Classification. For **RadChest-CT** and **MultiCenter-CT**, we first obtain lung masks using a publicly available pre-trained U-Net for lung segmentation [7]. We then extract the lung bounding box, resize the cropped volume to $48 \times 256 \times 256$, and normalize CT intensities for standardization. For **UCSF-MRI**, we directly resize each volume to $48 \times 256 \times 256$ and apply min–max normalization. Although resizing may introduce geometric distortion, LAD is designed to model and adapt to the resulting anisotropy. We train all classification models with a batch size of 16 using cross-entropy loss.

Hardware and Software. Experiments are implemented in PyTorch 2.6 (Python 3.12) and conducted on a single NVIDIA A800 (80GB) GPU. We use the AdamW optimizer for all tasks.

4 Results and Discussion

4.1 Main Results

Table 1 demonstrates the consistent performance gains of LAD across volumetric segmentation and diverse classification benchmarks. For the standard UNETR, LAD increases the mDice from 65.93% to 71.58% while reducing boundary errors (mHD95). In classification, LAD compensates for the ViT baseline's lack of spatial inductive bias, improving the RadChest-CT (T3) AUC from 0.498 to 0.678. Notably, LAD yields additional gains when integrated into stronger architectures like SwinUNETR and Primus. Crucially, LAD provides improved stability in limited-data regimes; it improves the EVA-02 baseline on the UCSF-MRI task, elevating the AUC from 0.347 to 0.584.

Static vs. Dynamic Scaling. To evaluate LAD against existing anisotropy-aware methods, we integrate AFPE [9] into the UNETR backbone, following the

Table 1. Comprehensive performance of LAD across segmentation and classification tasks. AFPE employs dataset-specific fixed scaling factors.

Seg./Cls. Model	Segmentation		Classification (AUC ↑)				
	mDice ↑	mHD95 ↓	T1	T2	T3	T4	T5
UNETR/ViT	65.93	50.21	0.609	0.542	0.498	0.519	0.678
+ AFPE	66.80	49.68	0.659	0.600	0.646	0.516	0.738
+ LAD (Ours)	**71.58**	**41.37**	**0.676**	**0.616**	**0.678**	**0.586**	**0.762**
SwinUNETR/Swin Trans.	72.83	41.79	0.657	0.602	0.625	0.532	0.769
+ LAD (Ours)	**73.43**	**41.04**	**0.693**	**0.631**	**0.636**	**0.541**	**0.813**
Primus/EVA-02	73.46	35.69	0.636	0.561	0.524	0.347	0.686
+ LAD (Ours)	**74.10**	**34.84**	**0.644**	**0.598**	**0.581**	**0.584**	**0.770**

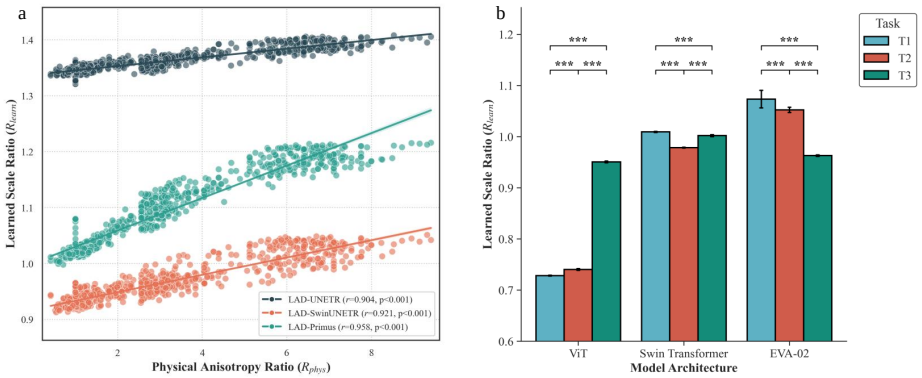

Fig. 2. Statistical analysis of geometry awareness and task adaptivity for LAD framework. We mathematically define the physical anisotropy ratio $R_{phys} = \frac{2 \cdot \text{Spacing}_Z}{\text{Spacing}_X + \text{Spacing}_Y}$ and the learned spatial scale ratio $R_{learn} = \frac{2 \cdot \text{Scale}_Z}{\text{Scale}_X + \text{Scale}_Y}$. **(a)** Correlation between R_{learn} and R_{phys} in volumetric segmentation, assessed using Pearson correlation analysis. **(b)** Task-adaptive distributions of R_{learn} across three RadChest-CT classification targets, specifically atelectasis for T1, lung opacity for T2, and bronchiectasis for T3. *** denotes $p < 0.001$ via Mann-Whitney U test.

original paper's focus on ViT-based architectures. Because the AFPE factor only reflects the dataset-level average, it fails to account for the inconsistent inter-scan spacing in AbdomenAtlas, resulting in limited segmentation gains compared to LAD's 71.58% mDice. While a static approach can benefit specific tasks like MultiCenter-CT, its rigidity reduces the UCSF-MRI AUC to 0.516. In contrast, LAD infers instance-level scales directly from image features, ensuring superior and consistent performance across all evaluated protocols.

4.2 Statistical Analysis for LAD Framework

Geometry-Spacing Correlation. To test whether LAD perceives physical dimensions, we compare R_{learn} with R_{phys} on the AbdomenAtlas test set. As shown in Fig. 2a, R_{learn} has a strong positive correlation with R_{phys} across all

Table 2. Ablation study across all tasks.

Method	Segmentation		Classification (AUC ↑)				
	mDice ↑	mHD95 ↓	T1	T2	T3	T4	T5
Baselines & Ablation:							
(1) Baseline (Standard)	65.93	50.21	0.609	0.542	0.498	0.519	0.678
(2) + RoPE (Isotropic)	69.26	45.25	0.643	0.581	0.578	0.463	0.742
(3) + RoPE (Anisotropic)	70.43	43.19	0.666	0.587	0.592	0.570	0.756
(4) **LAD (Full)**	**71.58**	**41.37**	**0.676**	**0.616**	**0.678**	**0.586**	**0.762**

backbones ($r = 0.904$–$0.958, p < 0.001, n = 1039$). This indicates that, under native acquisition resolutions, LAD captures the trend of physical anisotropy directly from image content, without relying on spacing metadata.

Task-Dependent Scale Adaptation. In contrast, on RadChest-CT where all images are resized to a fixed grid, Fig. 2b shows that R_{learn} values adapt significantly between different tasks ($p < 0.001$). Because the resizing process removes physical spacing differences, these shifts indicate that the model adjusts its scaling based on disease-specific imaging characteristics.

Together, these results indicate that the learned scales align with true physical spacing in segmentation and adapt to disease-specific imaging characteristics in classification, supporting both geometric fidelity and task-dependent adaptivity.

4.3 Ablation Study

We conduct an incremental ablation study using the UNETR(ViT) backbone to validate each proposed component, as detailed in Table 2.

Impact of Isotropic Assumptions. Adding standard Isotropic RoPE improves segmentation from 65.93% to 69.26% mDice and benefits several tasks. However, its rigid isotropic coordinate assumption limits instance-specific adaptability, leading to inconsistent performance across benchmarks. This limitation becomes evident in limited-data scenarios such as T4 (UCSF-MRI), where enforcing fixed isotropic geometry on resized volumes misaligns spatial representations and reduces the AUC from 0.519 to 0.463. In contrast, our Anisotropic RoPE allows the coordinate space to adapt to learned geometric scales, recovering T4 performance to 0.570 and further increasing segmentation to 70.43% mDice.

Necessity of Amplitude Modulation. While Anisotropic RoPE calibrates spatial phase, semantic recalibration along biased axes remains necessary. Integrating Adaptive Anisotropic Feature Modulation completes the full LAD framework, increasing segmentation to 71.58% mDice and notably enhancing performance on tasks like T3, which increases from 0.592 to 0.678. This dual-calibration of geometric phase and semantic amplitude achieves the most stable performance across both dense segmentation and global classification targets.

5 Conclusion and Outlook

This work introduces LAD, an architectural extension for 3D Transformers designed to adapt to anisotropic data by learning spatial priors. Across various segmentation and classification tasks, incorporating LAD into standard or foundation-style backbones yields consistent performance improvements. Beyond these results, LAD offers interpretability: the learned scale ratio can correlate with physical spacing in segmentation and adjust to imaging characteristics in classification. These findings suggest that adaptive spatial scaling helps handle data heterogeneity in 3D medical imaging. Future research may explore the integration of these adaptive mechanisms into the pre-training phase of 3D medical foundation models. Such an approach could facilitate the development of robust, resolution-agnostic representations across diverse large-scale datasets.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62576022), the National Key R&D Program of China under Grant No. 2023YFC2506502, Beijing Natural Science Foundation (L232131), and the Fundamental Research Funds for the Central Universities.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. https://doi.org/10.48550/arxiv.2404.00578
2. Blankemeier, L., Cohen, J.P., Kumar, A., Veen, D.V., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., Bluethgen, C., Jensen, M.E.K., Ostmeier, S., Varma, M., Valanarasu, J.M.J., Fang, Z., Huo, Z., Nabulsi, Z., Ardila, D., Weng, W.H., Junior, E.A., Ahuja, N., Fries, J., Shah, N.H., Johnston, A., Boutin, R.D., Wentland, A., Langlotz, C.P., Hom, J., Gatidis, S., Chaudhari, A.S.: Merlin: A vision language foundation model for 3d computed tomography. https://doi.org/10.48550/arXiv.2406.06512
3. Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P.: The university of california san francisco preoperative diffuse glioma MRI dataset. Radiology: Artificial Intelligence **4**(6), e220058 (2022). https://doi.org/10.1148/ryai.220058
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021). https://doi.org/10.48550/arXiv.2010.11929
5. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical Image Analysis **67**, 101857 (2021). https://doi.org/10.1016/j.media.2020.101857

6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 574–584 (2022). https://doi.org/10.1109/WACV51458.2022.00181
7. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem **4**(1), 50. https://doi.org/10.1186/s41747-020-00173-2
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
9. Jabareen, N., Yuan, D., Liu, D., Ten, F.W., Lukassen, S.: Anisotropic fourier features for positional encoding in medical imaging. In: Shape in Medical Imaging (ShapeMI) at MICCAI 2025. pp. 1–13. Springer, Daejeon, Republic of Korea (2025). https://doi.org/10.1007/978-3-032-06774-6_1
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
11. Mackin, D., Fave, X., Zhang, L., Yang, J., Jones, A.K., Ng, C.S., Court, L.: Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLOS ONE **12**(9), e0178524 (2017). https://doi.org/10.1371/journal.pone.0178524
12. Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D.P., Kainz, B., Glocker, B., Rueckert, D.: Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging **37**(2), 384–395 (2018). https://doi.org/10.1109/TMI.2017.2743464
13. Qu, C., Zhang, T., Qiao, H., Tang, Y., Yuille, A.L., Zhou, Z.: Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. Advances in Neural Information Processing Systems **36** (2023). https://doi.org/10.5555/3666122.3667714
14. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024). https://doi.org/10.1016/j.neucom.2023.127063
15. Wald, T., Roy, S., Isensee, F., Ulrich, C., Ziegler, S., Trofimova, D., Stock, R., Baumgartner, M., Köhler, G., Maier-Hein, K.: Primus: Enforcing attention usage for 3d medical image segmentation (2025), https://arxiv.org/abs/2503.01835
16. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images (2024). https://doi.org/10.48550/arXiv.2310.15161
17. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nature Communications **16**(1), 7866 (2025). https://doi.org/10.1038/s41467-025-62385-7